ERIK RUIST

Hantera tvärsnittsdata varsamt!

Under senare år har ekonomer intresserat sig alltmer för mikrodata. Makroekonomiska modeller har i viss mån ersatts av analyser på individnivå. Detta kan avspegla ett ökande intresse för individer och fördelningen mellan individer, men också att förändringar i makroekonomiska storheter visat sig kräva förklaringar på mikronivå: olika grupper av individer reagerar på olika sätt, och detta kan inte fångas av samband mellan makrostorheter. Samtidigt har tillgängligheten på mikrodata ökat, bl a genom lättheten att behandla stora datamängder och att samköra olika registermaterial.

Tillgängliga data uppfyller dock inte alla önskemål som användarna har. För många analyser av det dynamiska skeendet skulle man helst vilja följa varje individ under en längre period eller åtminstone vid två observationstillfällen. Sådana paneldata är dock tämligen ovanliga, av flera skäl. Det är därför oftast nödvändigt att basera analysen på en jämförelse mellan tvärsnitt vid två (eller flera) tidpunkter. Data vid varje tidpunkt består då av en mängd variabler för olika individer, som kan grupperas antingen efter någon kvalitativ egenskap eller efter storleken av någon variabel. Vanliga exempel är inkomstfördelningar och beräkningar av medelinkomsten i olika typer av hushåll.

En karakteristisk egenskap hos populationer av det slag som här studeras är att de är öppna, d v s individer strömmar inte bara mellan klasserna utan också ut och in i populationen. De individer som finns i en klass vid ett tillfälle är alltså inte desamma som de som fanns där vid föregående observation. Detta ställer till besvär, eftersom det gör att förändringar i t ex gruppernas medelinkomster mellan de två tillfälllena blir betydligt mer svårtolkade än de skulle ha varit i en statisk population eller om man haft paneldata.

Ett ofta citerat exempel på detta är inkomstutvecklingen för pensionärer. Medelinkomsten inom pensionärsgruppen har under det senaste decenniet stigit ganska rejält, vilket naturligtvis inte utan vidare kan tolkas så, att de som var pensionärer för tio år sedan haft stora inkomstökningar. Gruppens sammansättning har under tiden förändrats. Genom att nytillträdande pensionärer haft betydligt högre pension än de som avlidit under perioden har gruppens medelinkomst ökat.

Ändå presenteras ofta resultaten av undersökningarna på ett sätt som åtminstone av flertalet läsare torde tolkas som om de grundar sig på paneldata. Här ett exempel från pressreferatet av en nyligen offentliggjord SCB-undersökning av inkomstförändringar 1989-95 (SvD 1997-10-20):

För de flesta ålderspensionärer har perioden 1989-95 betytt små ökningar av inkomsten – men det har i alla fall varit ökningar. För de yngre pensionärerna, de mellan 65 och 74 år, steg snittinkomsten med 2,2 procent. För dem som är 75 år och äldre steg snittinkomsten med 8,5 procent.

Det är denna typ av analys och presentation jag vill ta upp till diskussion. I det citerade fallet förelåg väl ingen egentlig frågeställning. Avsikten var förmodligen ren deskription. Men ofta avser liknande analyser att undersöka hur någon ekonomisk-politisk åtgärd påverkat olika grupper i samhället. Då kan man tänka sig två olika typer av frågor som man kan ställa.

ERIK RUIST är professor emeritus i ekonomisk statistik vid Handelshögskolan i Stockholm.

Den för mig mest naturliga frågan är den som bäst skulle besvaras med hjälp av paneldata: Hur påverkades de som när åtgärden trädde i kraft var tvåbarnsföräldrar, pensionärer, arbetslösa etc? Eftersom man i allmänhet inte har tillgång till paneldata, får man försöka få ut så mycket information ur tvärsnittsdata som möjligt. Här finns ju en del metoder att tillgå, men det verkar som om de vore helt bortglömda vid det här laget.

Den andra typen av fråga har jag svårare att formulera, men kan väl exemplifieras så: "Har småbarnsföräldrar (pensionärer, studerande ungdom etc) det bättre ställt än motsvarande grupp förr i världen?" Denna fråga besvaras givetvis av data som visar bruttoförändringar för de olika grupperna. Men även här förefaller det som om det vore intressant att komplettera bruttosiffran med en undersökning av om förändringen delvis beror på åldersförskjutningar, regionala flyttningar el dyl.

Tyvärr är det sällan som det anges vilken typ av fråga som en analys skall svara på. Därför blir läsaren lätt vilseledd, särskilt om formuleringarna är försåtliga, medvetet eller omedvetet.

Effekter av politiska åtgärder

Under senare år har många analyser gjorts för att belysa hur olika ekonomiskpolitiska åtgärder påverkat olika grupper av individer eller hushåll. Bland annat har skattereformen 1991 analyserats på detta sätt. Jag skall emellertid välja ett något färskare exempel, där jag tycker att det borde vara möjligt att fördjupa analysen.

Den fördelningspolitiska redogörelsen i årets budgetproposition (Prop 1997/98:1 Bilaga 7) analyserar effekterna på olika typer av individer och hushåll av det ekonomiska saneringsprogrammet. Där görs först en s k regelanalys. Den kan sägas vara ett försök att skapa paneldata genom att utgå från inkomstuppgifterna för ett år innan åtgärderna vidtogs och för varje hushåll räkna fram hur det skulle drabbas av de olika skatte-, avgifts- och bidragsförändringarna. Resultaten jämförs sedan för olika grupper. Detta är uppenbarligen ett försök att svara på frågor av den första typen ovan. Det konstateras emellertid i motsvarande studie av skattereformen (Eklind m fl [1995]), att en regelanalys med nödvändighet blir statisk, d v s den tar ingen hänsyn till de eventuella beteendeförändringar som de olika åtgärderna medfört.

För att kunna inkludera effekterna av sådana förändringar måste man uppenbarligen jämföra en inkomstfördelning för ett år före åtgärderna med en efter desamma. Problemet är då att så mycket annat händer samtidigt att det är svårt att identifiera effekterna av just den åtgärd man vill studera.

I den fördelningspolitiska redogörelsen analyseras sålunda den totala förändringen i inkomstfördelningen mellan 1991 och 1997 (den senare erhållen som en framskrivning från 1995), alltså resultatet av alla förändringar som ägt rum under perioden, inklusive anpassningar till det nya regelverket. Redovisningen avser huvudsakligen förändringar för sådana grupper som heltidsarbetande, deltidare, företagare, arbetslösa, pensionärer och studerande. Bland de heltidsarbetande skiljer man på tre nivåer av arbetsinkomst: höginkomsttagare (den högsta decilen), medelinkomsttagare (decilerna 2–9) och låginkomsttagare (den lägsta decilen).

Resultaten redovisas bl a i form av förändringar av olika gruppers genomsnittliga inkomster:

- "Höginkomsttagarnas disponibla inkomster har sjunkit med ca 20 000 kronor eller ca 7 procent."
- "Ungdomar/studerande har fått den största minskningen av de disponibla inkomsterna."

Det är svårt att inte tolka detta som utsagor om inkomstutvecklingen för en given grupp av personer. Samtidigt är det uppenbart för den som tänker efter att alla de ovan uppräknade grupperna är öppna, d v s det sker en ständig ut- och inströmning av individer i dem. När man jämför medelinkomsten inom en grupp vid två olika tillfällen är det alltså inte samma individer som jämförs. Frågan är då vad jämförelserna egentligen innebär. Det framgår inte klart om frågeställningen är av typ 1 eller 2 ovan. Något försök görs inte heller att analysera effekten av störande variabler på jämförelsen. Det skulle ju annars vara lätt att korstabulera de analyserade klasserna mot ålder, arbetslöshet och andra variabler som kan misstänkas snedvrida jämförelserna. Förr användes ju också s k standardisering, där cellerna i båda populationerna sammanvägdes med samma fördelning över de kritiska variablerna för att eliminera effekter av förskjutningar. Det vore inte fel att göra så idag också!

I den här refererade *Fördelningspolitiska redogörelsen* är författarna utan tvekan medvetna om svårigheterna och påpekar bl a att "gruppernas sammansättning har förändrats under perioden". Att i den analyserande texten lägga in ett antal reservationer av innebörden att andra än de påtalade orsakerna också kan ha inverkat, är emellertid som alla vet helt verkningslöst. En politiker eller annan användare som tycker sig ha fått bekräftelse på sin älsklingstes glömmer – eller gömmer undan – reservationerna.

Det är ju också så, att skillnaden mellan bruttoutvecklingen för en viss grupp och den man skulle ha fått, om paneldata varit tillgängliga, kan vara avsevärd. Detta är ett klassiskt problem i statistiken, numera kanske oftast karakteriserat som effekten av störande variabler. För att ändå påminna om hur effekten kan uppkomma har jag konstruerat ett exempel, som medvetet är mycket förenklat.

Enkelt exempel

Låt oss betrakta en ort där arbetsmarknaden för heltidsarbetande består av ett enda företag. Det är ett stabilt företag med 45 anställda. Varje år anställs en 20-åring, som får en månadslön på 10 000 kr (räknat i något lämpligt års penningvärde). Varje år får alla anställda ett reallönelyft - ett ålderstillägg - på 3 procent, så att den äldste i företaget, 64-åringen, har en månadslön på 36 715 kr. Varje år avgår en 65-åring med pension.

Betraktar vi det här samhället år från år, kan vi konstatera att för alla grupper (låginkomsttagare, medelålders anställda etc), hur de än definieras, är den genomsnittliga inkomstnivån (realt) konstant. Samtidigt har varje individ en jämn inkomstökning med 3 procent om året ända till pensionsåldern, vilket skulle synas i paneldata.

Vi antar nu att företaget drabbas av en efterfrågeminskning och anser sig behöva skära ner arbetsstyrkan med fem personer. Man överväger två alternativ för detta: antingen förtidspensioneras de fem äldsta omedelbart, och pensionsåldern sänks till 60 år, eller också införs anställningsstopp under fem år.

Om någon av dessa åtgärder vidtas, kommer naturligtvis inkomststatistiken att påverkas. *Tabell 1* visar förändringen i medelinkomst för förvärvsarbetande "höginkomsttagare" (de översta 10 procenten), "medelinkomsttagare" (de mellersta 80 procenten) och "låginkomsttagare" (de understa 10 procenten) efter fem år enligt de båda alternativen, jämfört med det statiska utgångsläget.

Förändringarnas riktning är självklar: i det första fallet tar man bort de fem högsta inkomsterna ur populationen, i det senare fallet de fem lägsta.

Men om man nu bara har dessa makroindikatorer på vad som hänt med inkomstfördelningen, vilka slutsatser drar man då? Ja, tydligen har låginkomstgruppen drabbats minst, och i fallet nyrekryte-

Tabell 1 **Förändring av medelinkomsten över en femårsperiod för olika inkomsttagargrupper**

	Förtidspensionering	Nyrekryteringsstopp
	procentuell förändring från utgångsläget	
Höginkomsttagare	–13	+ 1
Medelinkomsttagare	– 8	+ 7
Låginkomsttagare	– 1	+15

ringsstopp rentav fått det väsentligt bättre! I själva verket är det ju så att fem personer med de lägsta inkomsterna har försvunnit ur populationen, varför medeltalet av de lägsta inkomsterna bland de kvarvarande blir högre. Genom att populationen systematiskt förändras blir förändringstalen för medeltalen meningslösa eller i varje fall mycket svårtolkade. Man kan knappast gissa att de individer som är kvar i populationen har helt oförändrade villkor, vilket ju är fallet.

Den ekonomiska verkligheten är naturligtvis mycket mer komplicerad än detta exempel, även om utan tvivel den här beskrivna mekanismen är en av många som påverkar de observerade medeltalen. Att det sedan finns andra orsaker gör ju inte att det är lättare att tolka innebörden av observerade förändringar av medeltal, sådana som i *Tabell 1*.

Det svåråtkomliga informationsinnehållet i jämförelser mellan grupper i två tvärsnitt av öppna populationer gör att jag är tveksam till nyttan av dem i de fall där frågeställningen är den första av de två som skisserades ovan. För att verkligen kunna få fram hur exempelvis inkomsten för olika grupper av individer utvecklat sig krävs givetvis paneldata. Tillgången på sådana är mycket begränsad, men att det ändå går att göra analyser av detta slag visas exempelvis av Uddhammar [1997]. Han beräknar bl a överströmningen mellan olika inkomstklasser över en sexårsperiod för olika åldrar, kön, socioekonomiska grupper etc. Tyvärr visar hans studie också en nackdel som vidlåder panelstudier: de är tidskrävande. De år som han jämför är 1985 och 1991. Om man är ute efter att analysera den senaste utvecklingen är en sådan eftersläpning naturligtvis förödande.

Det förefaller som om finansdepartementet och SCB har goda motiv för att starta ett gemensamt projekt med syfte att möjliggöra panelstudier av inkomstfördelningen, om än inte lika ambitiösa som Uddhammars.

Sammanfattningsvis är min uppmaning till ekonomer: var försiktig vid tolkning av denna typ av statistisk information. Försök åtminstone att genom korsklassificeringar eller standardberäkningar eliminera de mest uppenbara störningseffekterna. Förändringsprocesserna är emellertid så mångfasetterade att det är lätt att glömma (eller för läsaren att nonchalera) vissa förklaringsgrunder. Men – framför allt – uttryck resultaten så att läsaren inte kan misstolka dem!

Referenser

Eklind, B, Hussénius, J & Johansson, R, [1995], *Fördelningseffekter av skattereformen*, SOU 1995:104, bilaga 5

Regeringens proposition [1997/98, bilaga 7], *Fördelningspolitisk redogörelse*

Svenska Dagbladet, [1997], "Pensionärernas inkomster har ökat under 90-talet", 20 oktober 1997

Uddhammar, E, [1997], *Arbete, välfärd och bidrag*, City University Press, Stockholm.